Continual learning: the impact of task similarity

Neural Networks

Prof. Sebastian Goldt

Lorenzo Bardone
Ariel Surya Boiardi
Harshith Gowrachari

27/05/2022

SISSA

Introduction

Continual learning

Continual learning : Ability to learn many tasks in sequence

Major obstacle of continual learning :

- Catastrophic forgetting

Experimental setup

Teacher-Student setup

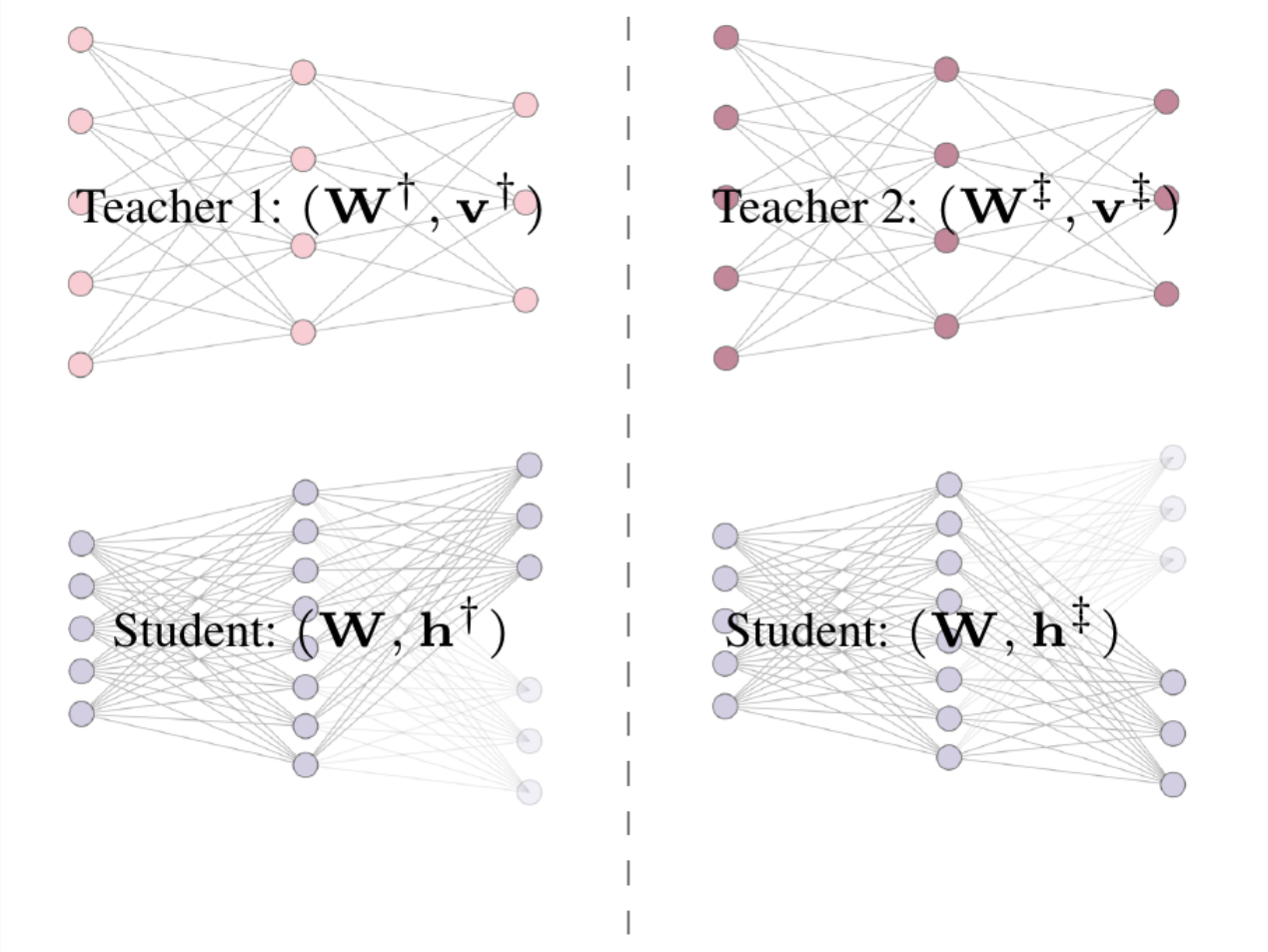


Figure 1: Teacher student setups for continual learning

Teacher-Student setup implementation

Teacher class :

- input dimension, D = 500
- one hidden unit, M = 1
- compute labels

Student class :

- Input dimension, D = 500
- Two hidden units, K = 2
- Compute labels

Generation of weights with given overlap

Starting from Gaussian i.i.d. vectors, we generate a horthonormal basis $\{\tilde{w}_1, \tilde{w}_2\}$ for an hyperplane in $\mathbb{R}^D$ by Grahm-Schmidt procedure, then fixed

$$\theta = \arccos(\mathit{overlap}),$$

weight vectors are taken to be

$$w_1 = \tilde{w}_1$$
$$w_2 = \cos(\theta)\tilde{w}_1 + \sin(\theta)\tilde{w}_2.$$

Experiments

Implementation

Initializations

- The teachers $\mathcal{T}_1$ and $\mathcal{T}_2$ are chosen with a fixed overlap of the first layer weights
- The weights of the student are from a Xavier Normal
- There is a unique test set with 10000 i.i.d $\sim \mathcal{N}(0, 1)$

Training

- Number of steps: 20000
- Solver: SGD
- Loss: Mean Squared Error
- Step size: 1/D=0.002
- Online learning regime: GD uses a new point $\sim \mathcal{N}(0, 1)$ at each step

Implementation

Algorithm structure

1. Initialise the two teachers $\mathcal{T}_1$, $\mathcal{T}_2$ and the student $\mathcal{S}$
2. For $i = 0, \ldots, 10000$ do:
   - Train $\mathcal{S}$ on $\mathcal{T}_1$
   - Test both tasks
3. Switch
4. For $i = 0, \ldots, 10000$ do:
   - Train $\mathcal{S}$ on $\mathcal{T}_2$
   - Test both tasks

Continual learning on identical tasks

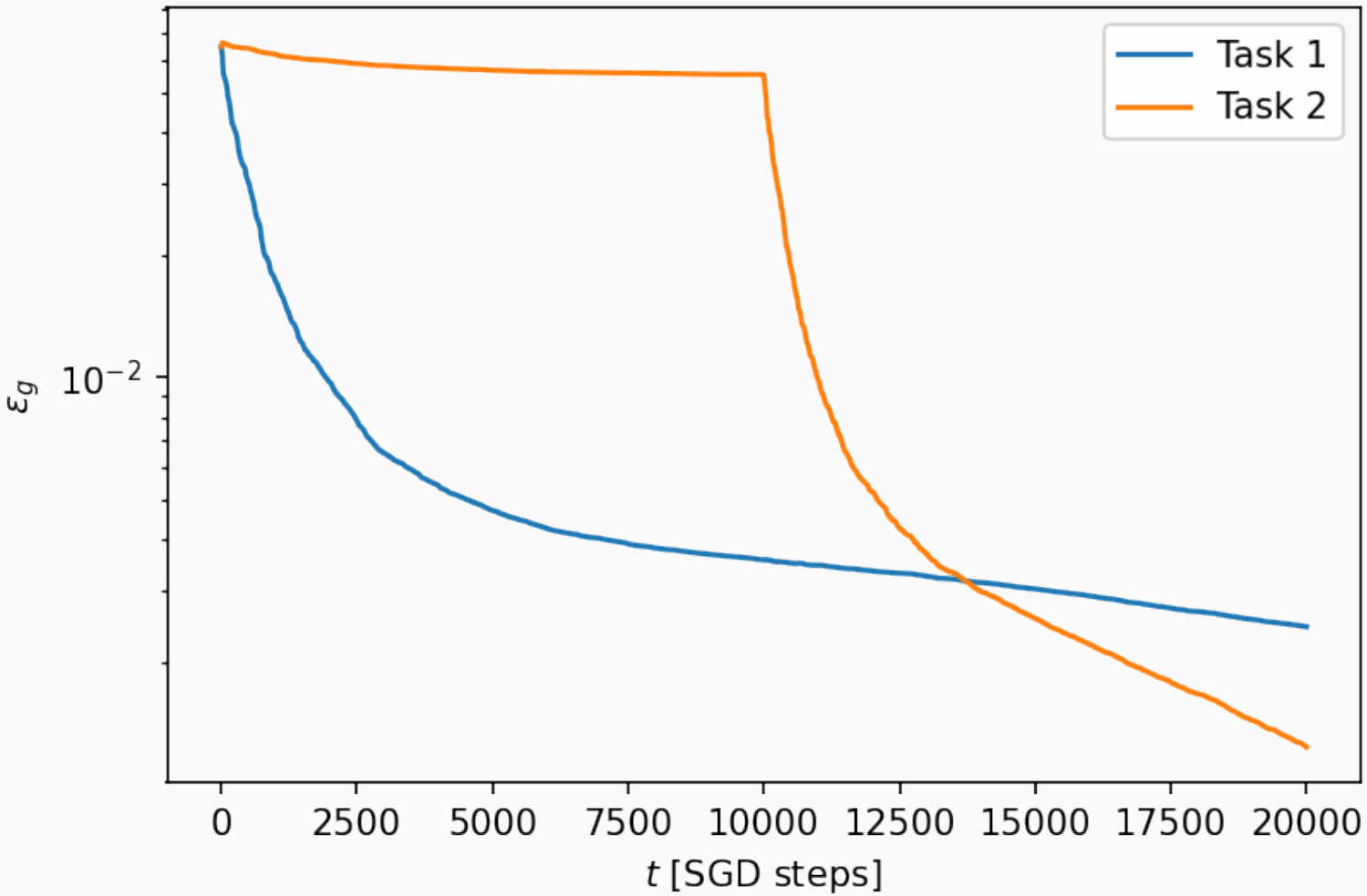


Figure 2: Training history with $\mathcal{T}_1 = \mathcal{T}_2$

Experiments

The impact of task similarity

Generalization error for varying task similarity

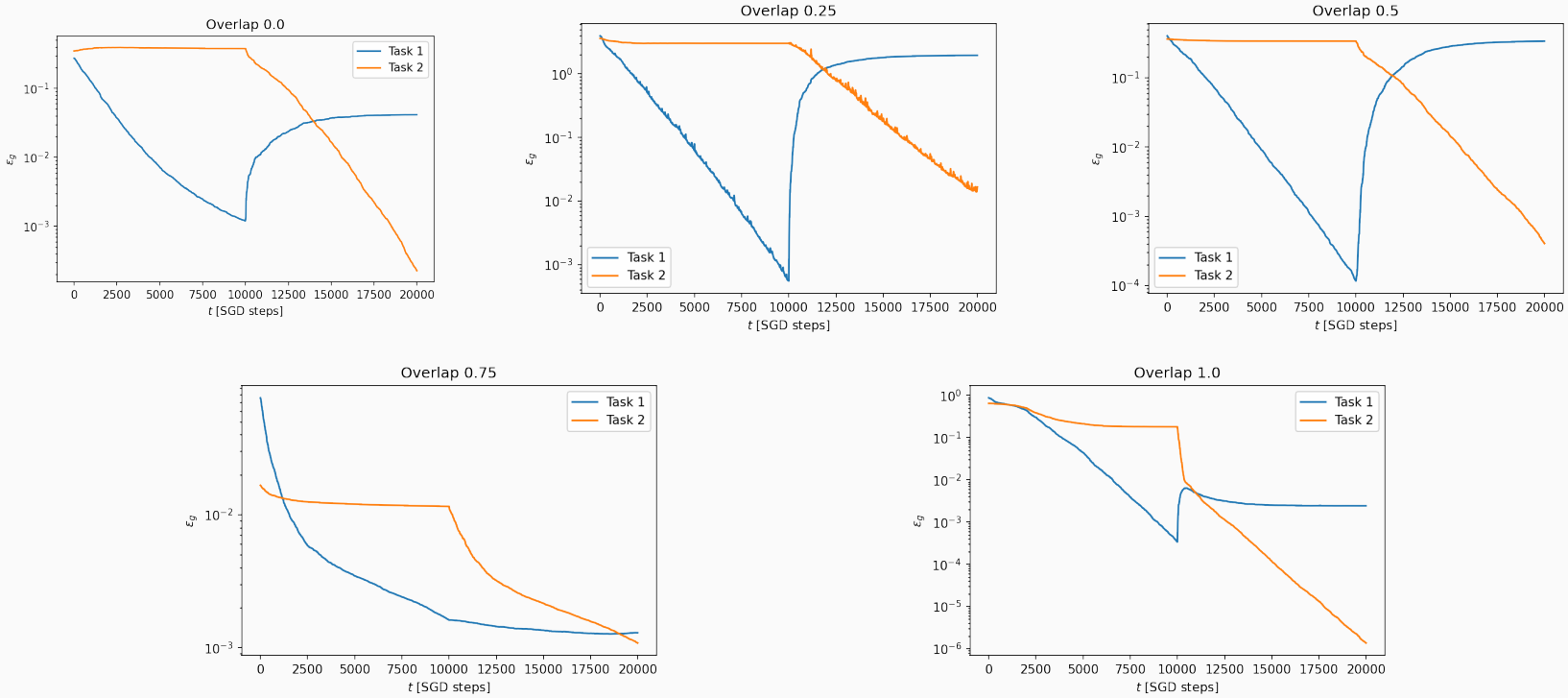


Figure 3: Continual learning is worst at intermediate task similarity. Experiment with 20×10^3 runs of online SGD.

Forgetting

Forgetting

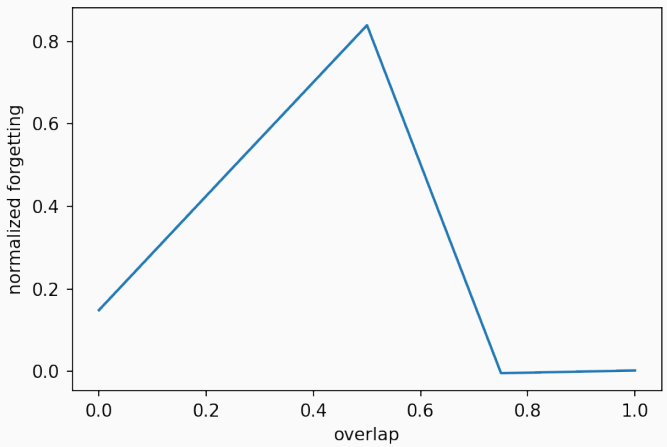
$$\mathcal{F} = \epsilon_g(t_2) - \epsilon_g(t_1);$$

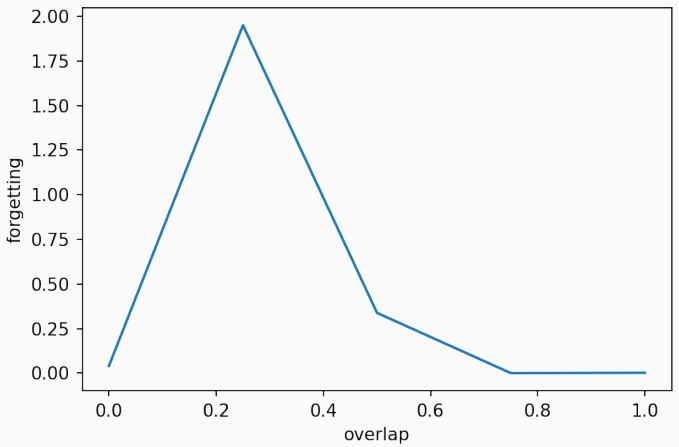
t_1 last step of training on task 1

t_2 last step of training on task 2

Normalized forgetting

$$\mathcal{F}_N = \frac{\mathcal{F}}{|\epsilon_g(0) - \epsilon_g(t_1)|}.$$

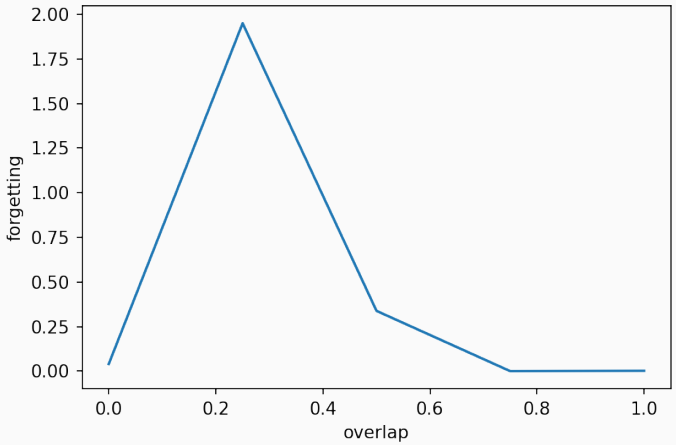




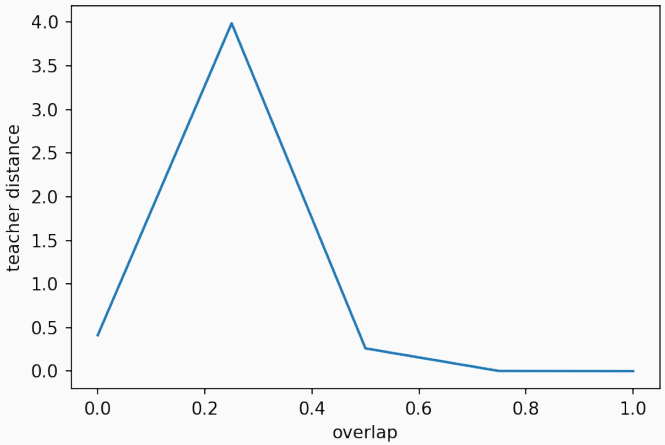
Teacher similarity on the dataset

Teacher similarity on data

$$\Delta_{dataset}(\mathcal{T}_1, \mathcal{T}_2) = \mathop{\mathbb{E}}_{x \in datased} \left[(\mathcal{T}_1(x) - \mathcal{T}_2(x))^2 \right]$$



(a) Forgetting

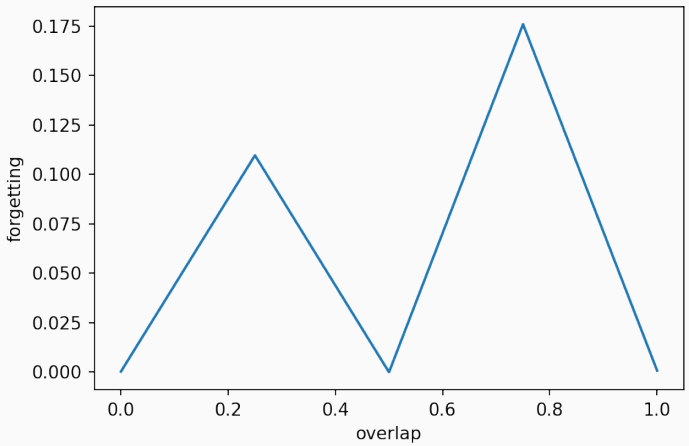


(b) Teacher distance

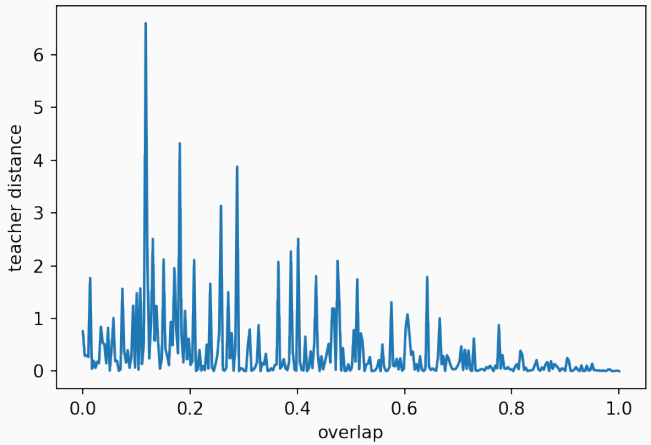
Discussion

Instability

- Heavy dependence on random initializations
- In our implementation overlap does not always predict teacher similarity



(a) Forgetting



(b) Teacher-Teacher similarity distribution

Experiments

Iterative adjustments

Continual *lifelong* learning

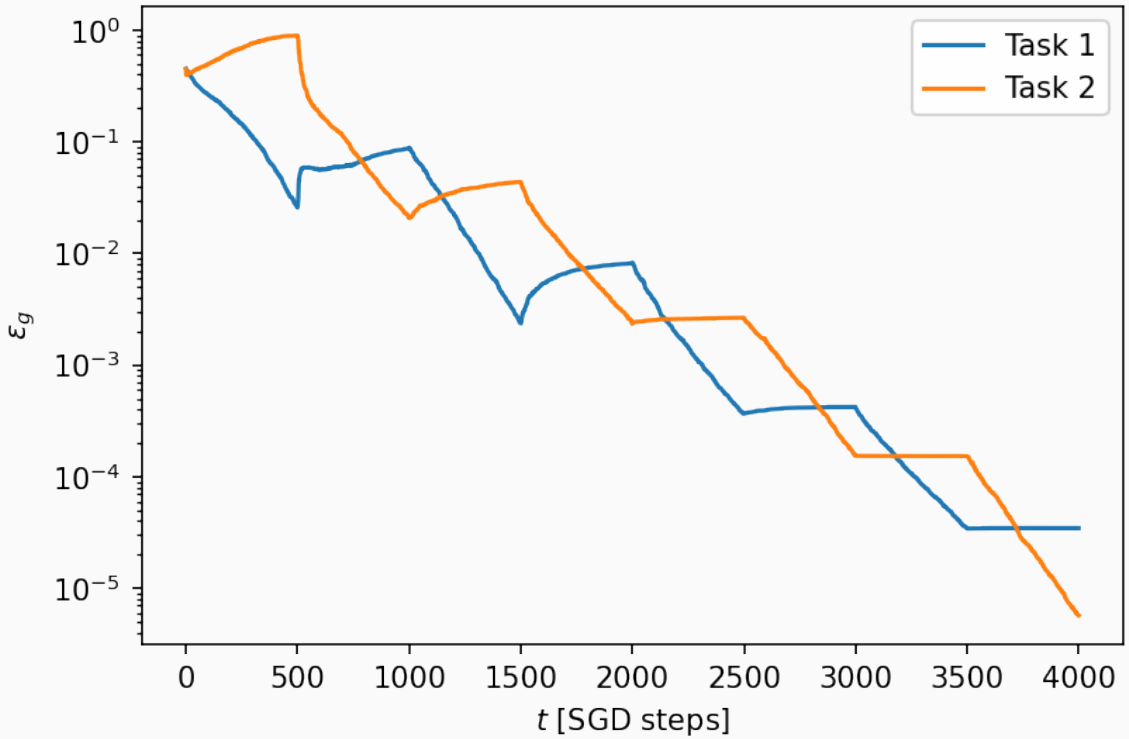
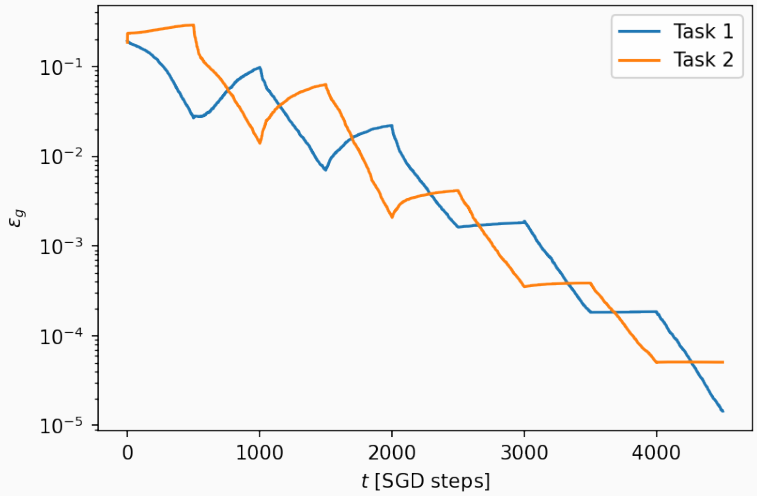


Figure 7: Continual learning with multiple task switch. Student trained on two teachers with overlap $\frac{1}{2}$ in 8 sessions of 5×10^3 SGD steps each.

Activation of student neurons



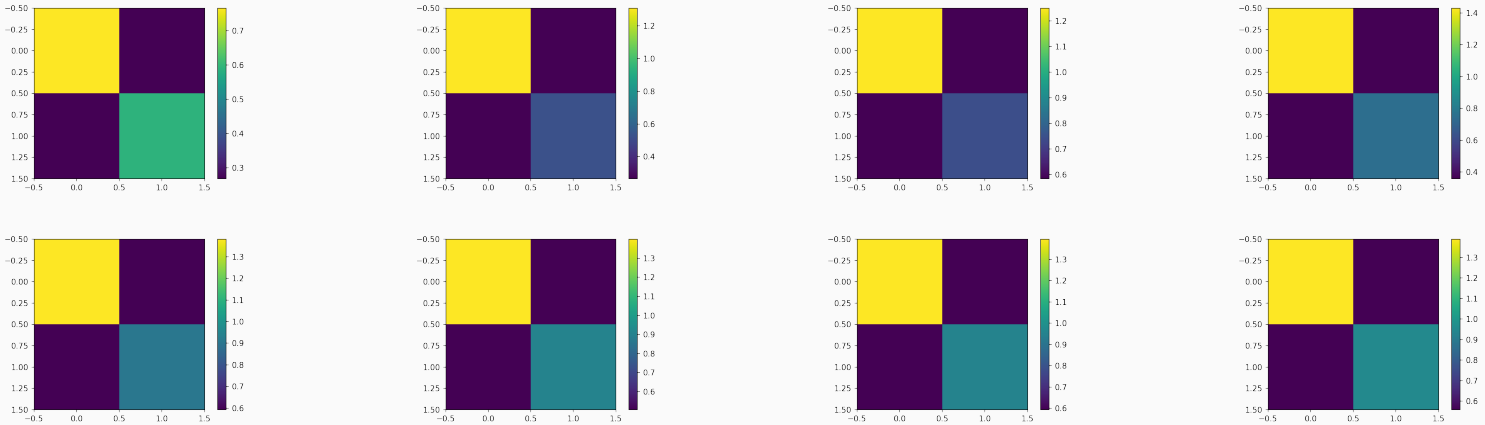
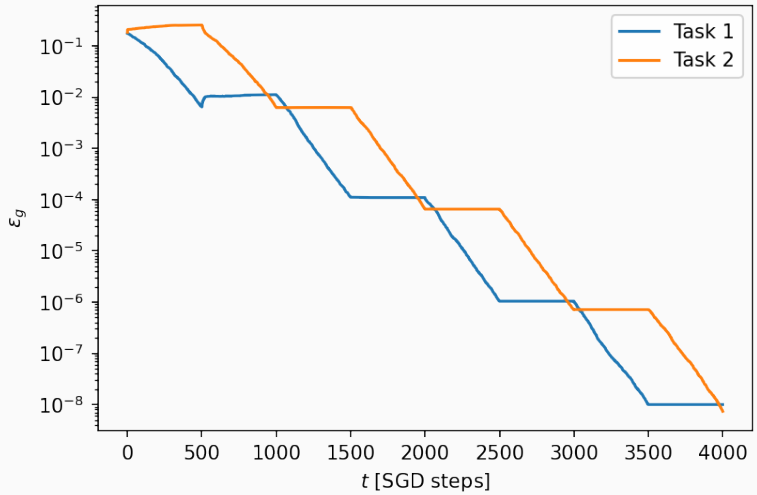


Figure 8: Student activation correlations $W^T W$

Activation of student neurons CONTINUED



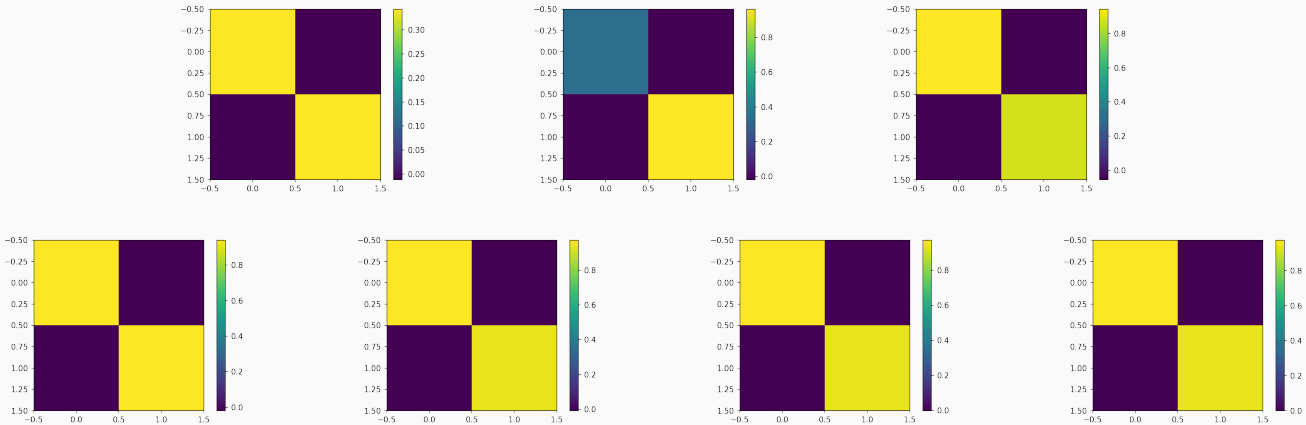


Figure 9: Student activation correlations $W^T W$

Experiments

Student with one head

Experimental setup

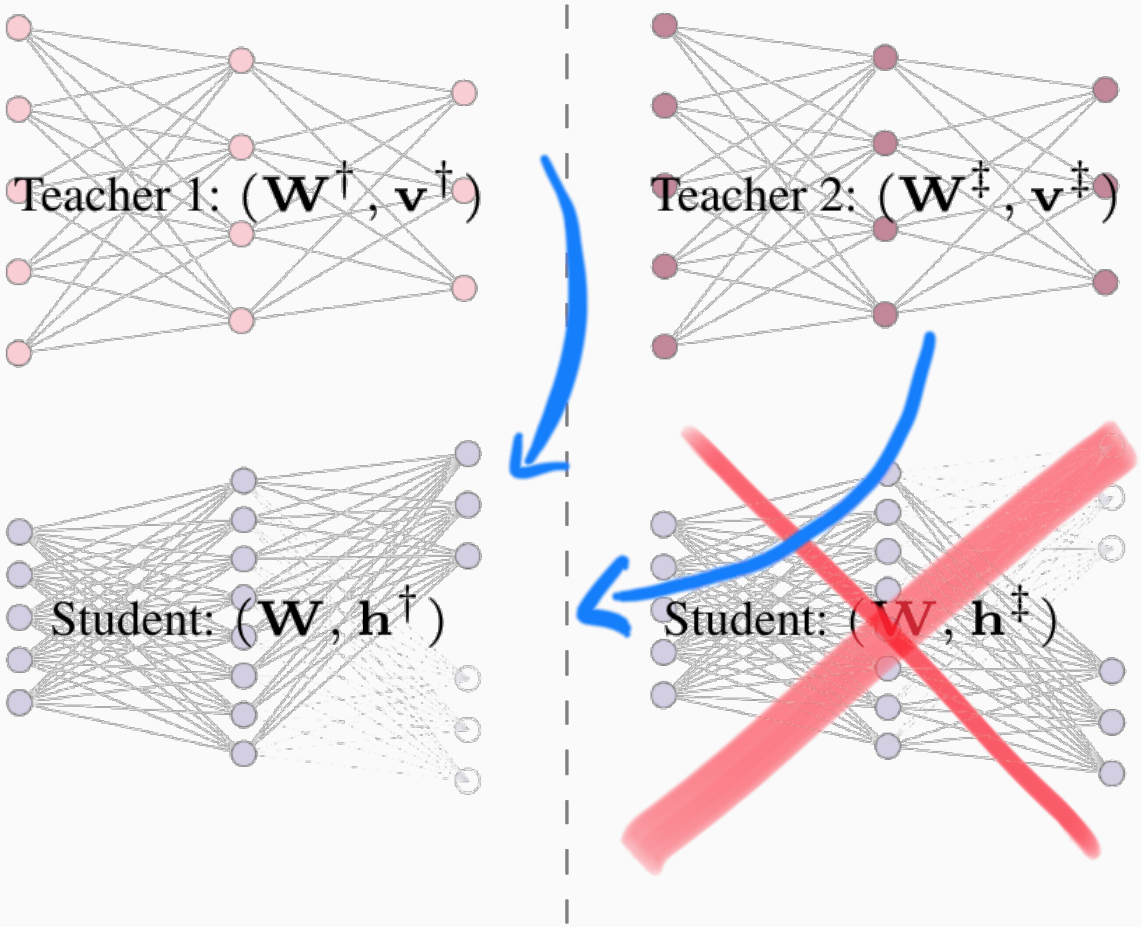


Figure 10: New teacher-student setup

Learning histories

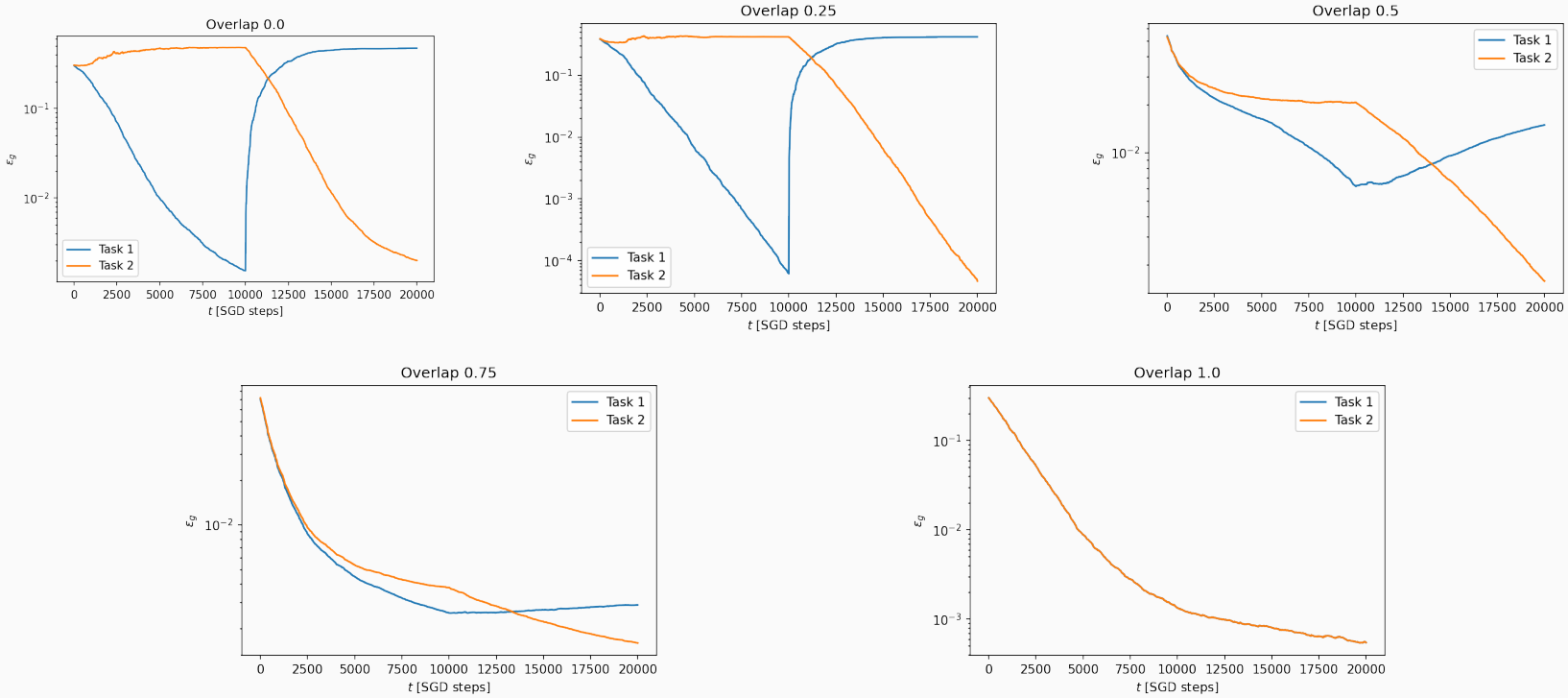


Figure 11: Continual learning improves for increasing task similarity. Experiment with 20×10^3 runs of online SGD.

Forgetting

Forgetting

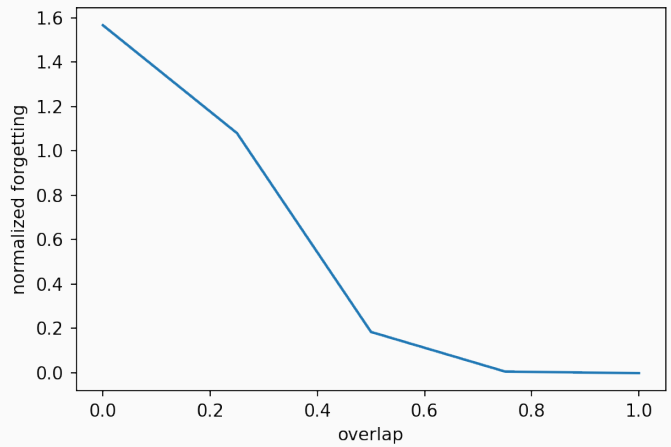
$$\mathcal{F} = \epsilon_g(t_2) - \epsilon_g(t_1);$$

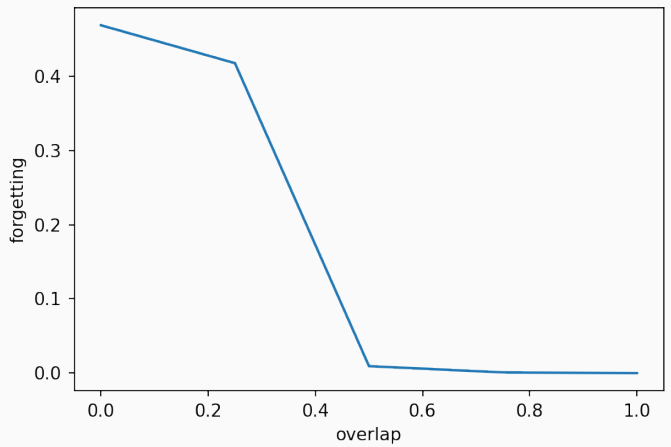
t_1 last step of training on task 1

t_2 last step of training on task 2

Normalized forgetting

$$\mathcal{F}_N = \frac{\mathcal{F}}{|\epsilon_g(0) - \epsilon_g(t_1)|}.$$

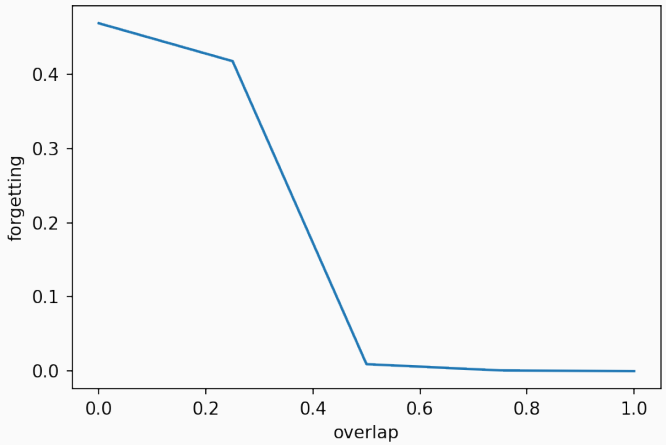




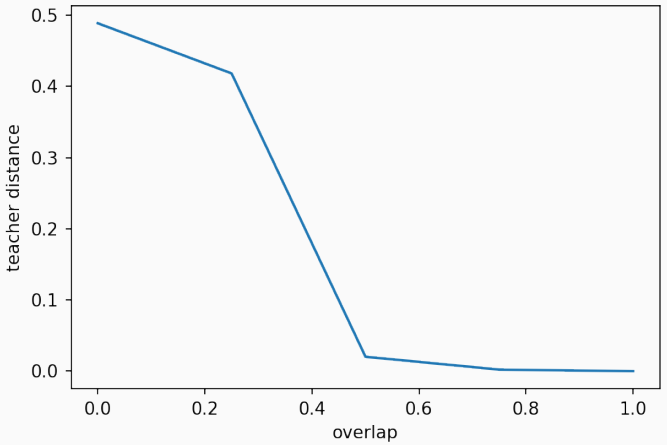
Teacher similarity on the dataset

Teacher similarity on data

$$\Delta_{dataset}(\mathcal{T}_1, \mathcal{T}_2) = \mathop{\mathbb{E}}_{x \in datased} \left[(\mathcal{T}_1(x) - \mathcal{T}_2(x))^2 \right]$$



(a) Forgetting



(b) Teacher distance

Thank you for your attention!